ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ассистент |  |  |  | В.В. Боженко |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ |
| Анализ связей между признаками двумерного набора данных |
| по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. № | 4917 |  |  |  | Д.В. Потемкин |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург 2022

1. **Цель работы**

изучение связи между признаками двумерного набора данных.

**Вариант 15:** 5и7gym\_churn

Информация о клиентах фитнес клуба:

1. Пол
2. Близкое расположение
3. Сотрудник компании партнера
4. По промо друзей
5. Указан ли телефон
6. Длительность текущего абонемента
7. Посещение групповых занятий
8. Возраст
9. Средние траты на дополнительные услуги
10. Количество месяцев до окончания абонемента
11. время с момента первого обращения в фитнес-центр (в месяцах)
12. Средняя частота посещений
13. Средняя частота посещений в текущем месяце
14. Покинул ли человек фитнес клуб
15. **Ход работы**

Импортируем датафрейм 5и7gym\_churn.csv и проведём предварительную обработку данных (Рисунок 1):

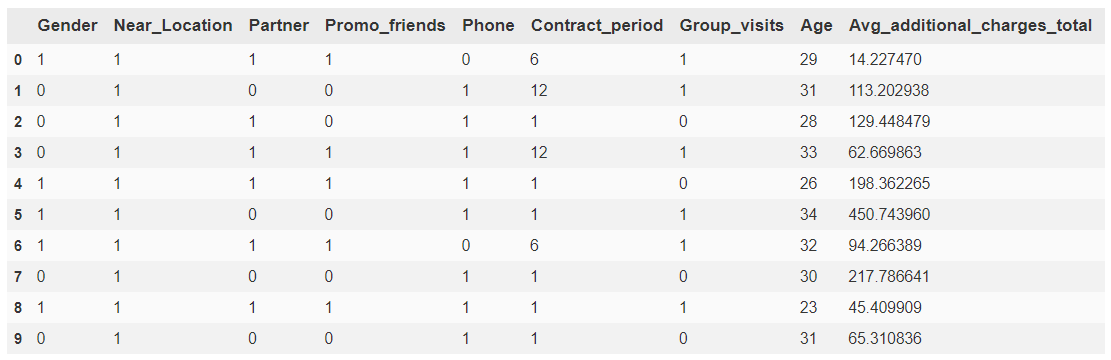


Рисунок 1 – Импортированный датафрейм

Построим точечные диаграммы (матрицы диаграмм рассеяния) (Рисунки 2-4):

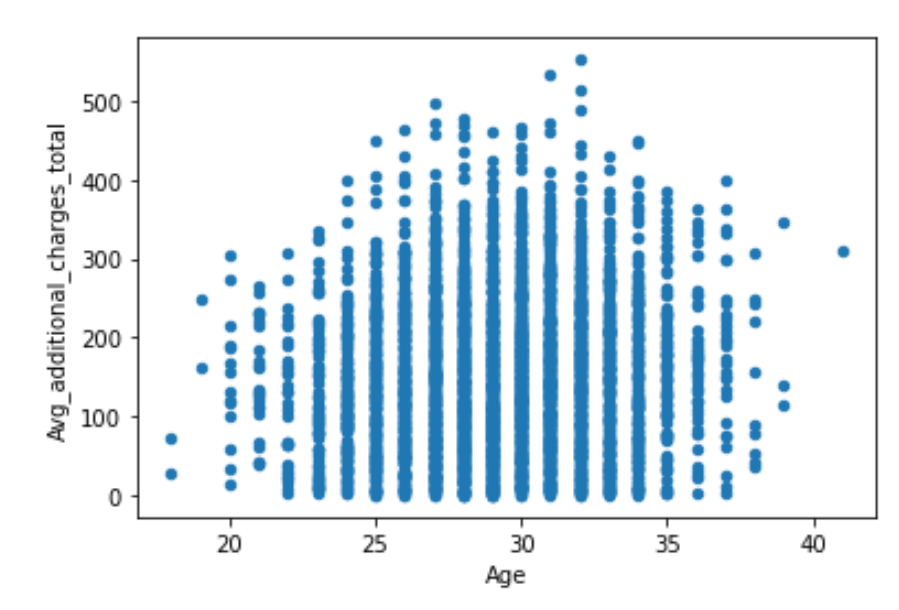


Рисунок 2 - Точечная диаграмма затрат на дополнительные услуги от возраста

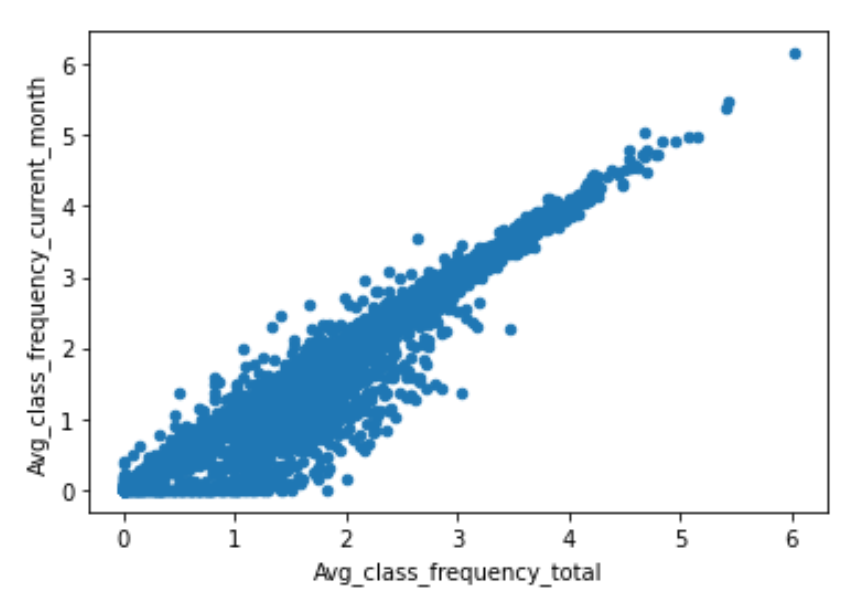


Рисунок 3 - Точечная диаграмма общей средней частоты посещений от средней частоты посещений в текущем месяце

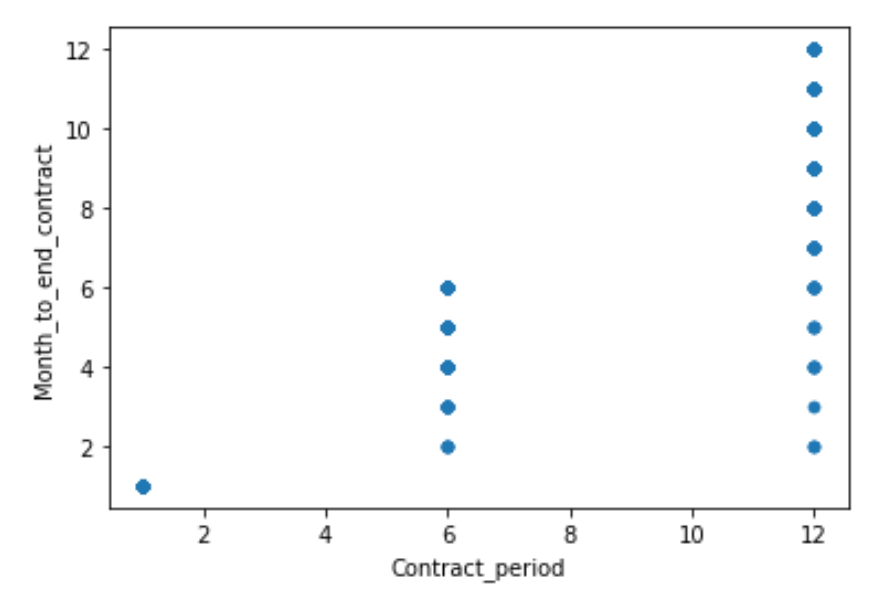


Рисунок 4 - Точечная диаграмма количества месяцев до окончания абонемента от длительности текущего абонемента

Построим матрицу диаграмм рассеяния (Рисунок 5):

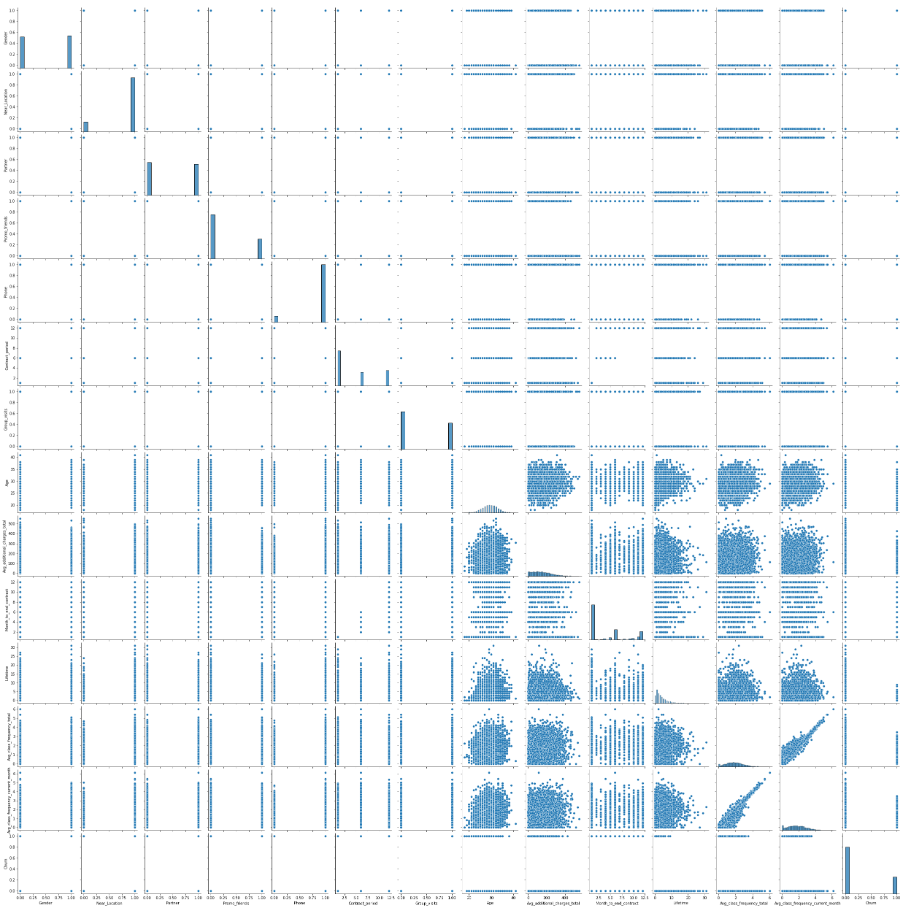


Рисунок 5 – Матрица диаграмм рассеивания

Из данной матрицы видно, что корреляция в данном датафрейме присутствует между двумя парами столбцов Avg\_class\_frequency\_total и Avg\_class\_frequency\_current\_month, и Contract\_period и Month\_to\_end\_contract

Вычислим коэффициент корреляции Пирсона для оценки связи между переменными (Рисунок 6):

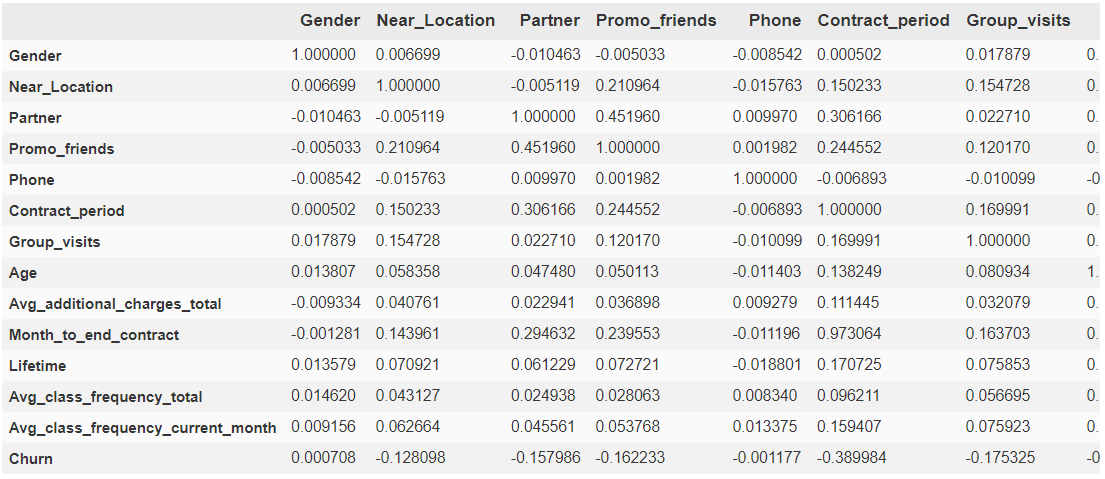


Рисунок 6 – таблица коэффициентов корреляции Пирсона

Наиболее больших значений (0.952739) и (0.973064) коэффициент корреляции Пирсона достигает между значениями Avg\_class\_frequency\_total и Avg\_class\_frequency\_current\_month, Contract\_period и Month\_to\_end\_contract и что подтверждает наличие взаимосвязи между значениями в этих парах.

Построим heatmap (тепловую карту корреляции) (Рисунок 7):

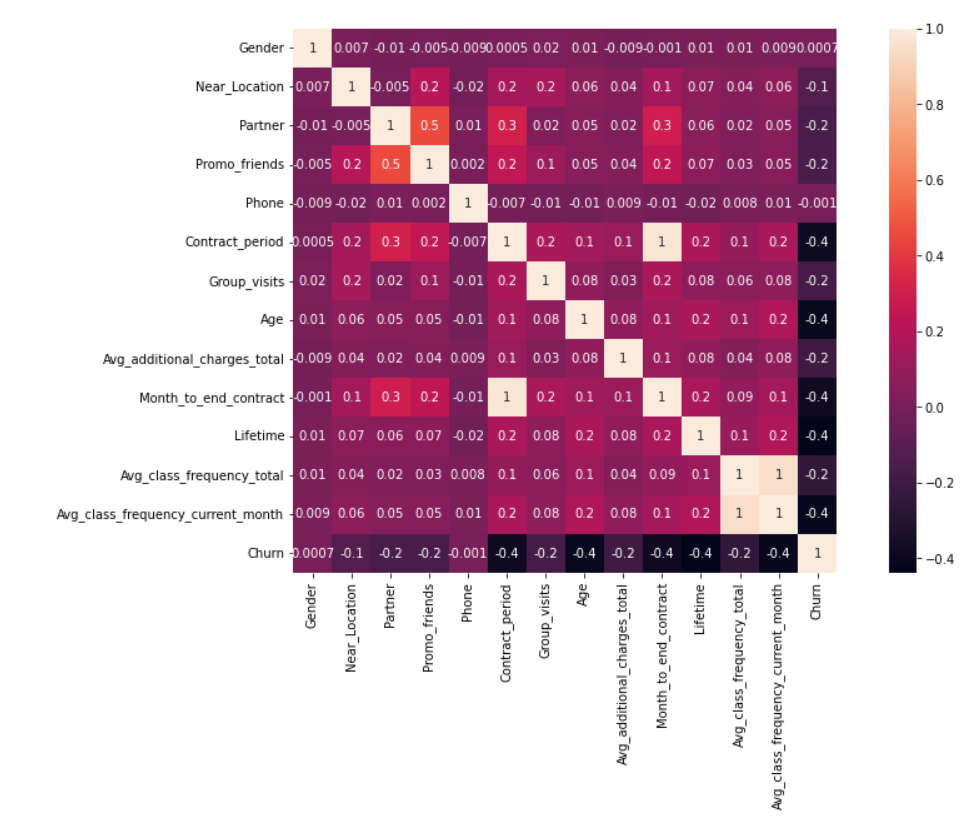


Рисунок 7 – Тепловая карта корреляции

Ссылка на GitHub с проектом:

1. **Вывод:**

в ходе выполнения лабораторной работы были изучены методы нахождения связей между признаками двумерного набора данных. Был импортирован и обработан датафрейм, содержащий информацию о посетителях фитнес клуба 5и7gym\_churn. На основе данного датасета были построены точечные диаграммы рассеяния, вычеслены коэффициенты корреляции Пирса между парами наборов значений и построена тепловая карта корреляции. В ходе исследования данных было выяснено, что в данном датафрейме связь присутствует между двумя парами столбцов Avg\_class\_frequency\_total (Общая средняя частота посещений) и Avg\_class\_frequency\_current\_month (Средняя частота посещений в текущем месяце), и Contract\_period (Длительность текущего абонемента) и Month\_to\_end\_contract (Количество месяцев до окончания абонемента).

**Дополнительное задание:**

Построить boxplot по любому столбцу, вывести describe, проанализировать, построить гистограмму (plot.bar) пришёл по промо друзей/не пришёл по промо друзей.

Построим Boxplot по столбцу 'Age' и выведем его описание при помощи describe (Рисунки 8-9):

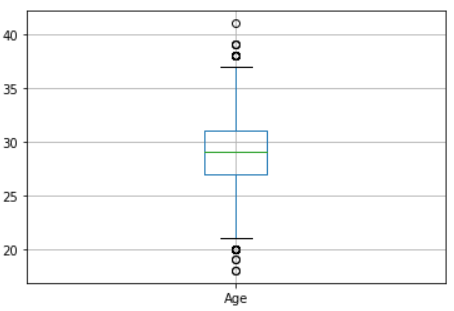


Рисунок 8 – Диаграмма размаха (Boxplot) по столбцу 'Age'

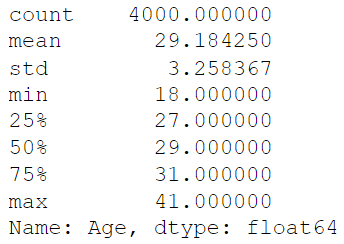


Рисунок 9 – Описание столбца 'Age'

Из диаграммы размаха видно, что большинство людей, посещающих спортзал имеют возраст от 21 до 37 лет, причём 50% из них - от 27 до 31 года. Однако, существует несколько выбросов: 18, 19, 20, 38, 39 и 41 год. возраст человека в центре ранжированного ряда составляет 29 лет. Из описания столбца видно, что среднее арифметическое возрастов, также очень близко к значению 29, а среднеквадратическое отклонение ≈ 3.26.

Построим гистограмму (Plot.bar) пришёл по промо друзей/не пришёл по промо друзей (Рисунок 10):

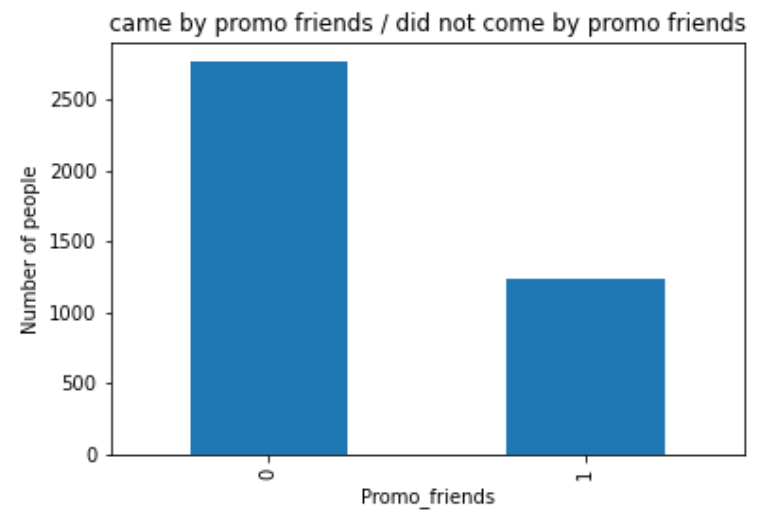


Рисунок 10 – Гистограмма по людям пришедшим и не пришедшим по промо друзей

Из гистограммы видно, что людей пришедших по промокоду друзей более чем в два раза меньше, нежели тех, кто записался в спортзал сам.